Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly

Thind, Anupriya Kaur ; Wicker, Thomas ; Šimková, Hana ; Fossati, Dario ; Moullet, Odile ; Brabant, Cécile ; Vrána, Jan ; Doležel, Jaroslav ; Krattinger, Simon G

Abstract: Cereal crops such as wheat and maize have large repeat-rich genomes that make cloning of individual genes challenging. Moreover, gene order and gene sequences often differ substantially between cultivars of the same crop species. A major bottleneck for gene cloning in cereals is the generation of high-quality sequence information from a cultivar of interest. In order to accelerate gene cloning from any cropping line, we report 'targeted chromosome-based cloning via long-range assembly' (TACCA). TACCA combines lossless genome-complexity reduction via chromosome flow sorting with Chicago long-range linkage to assemble complex genomes. We applied TACCA to produce a high-quality (N50 of 9.76 Mb) de novo chromosome assembly of the wheat line CH Campala Lr22a in only 4 months. Using this assembly we cloned the broad-spectrum Lr22a leaf-rust resistance gene, using molecular marker information and ethyl methanesulfonate (EMS) mutants, and found that Lr22a encodes an intracellular immune receptor homologous to the Arabidopsis thaliana RPM1 protein.

DOI: https://doi.org/10.1038/nbt.3877

Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly

Anupriya Kaur Thind[1+], Thomas Wicker[1+], Hana Šimková[2], Dario Fossati[3], Odile Moullet[3], Cécile Brabant[3], Jan Vrána[2], Jaroslav Doležel[2], Simon G. Krattinger[1*]

[1]**Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland**
[2]**Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic**
[3]**Institute for Plant Production Sciences, Agroscope, Switzerland**

[+]These authors contributed equally to this work
*Correspondence should be addressed to S.G.K. (skratt@botinst.uzh.ch)

Address for correspondence:

Department of Plant and Microbial Biology
University of Zurich
Zollikerstrasse 107
8008 Zurich, Switzerland
Phone: +41 44 634 82 33
Fax: +41 44 634 82 11

Cereal crops such as wheat and maize have large repeat-rich genomes that make cloning of individual genes challenging. Moreover, gene order and gene sequences often differ substantially between cultivars of the same crop species[1-4]. A major bottleneck for gene cloning in cereals is the generation of high-quality sequence information from the cultivar of interest. In order to accelerate gene cloning from any cropping line, we report 'targeted chromosome-based cloning via long-range assembly' (TACCA). TACCA combines lossless genome complexity reduction via chromosome flow sorting with Chicago long-range linkage[5] to assemble complex genomes. We applied TACCA to produce a high-quality (N50 of 9.76 Mb) *de novo* chromosome assembly of the wheat line 'CH Campala *Lr22a*' in only four months. Using this assembly we cloned the broad-spectrum *Lr22a* leaf-rust resistance gene using molecular marker information and EMS mutants and found that *Lr22a* encodes an intracellular immune receptor homologous to the Arabidopsis RPM1 protein.

While the world population continues to grow, the arable land per capita is decreasing[6]. To ensure food security, agriculture will require high-yielding crops that can withstand diseases, pests and adverse climatic conditions. A better understanding of genes that control these important traits may enable breeding of crop cultivars capable of feeding the 9-10 billion people expected by 2050.

'Positional cloning' or 'map-based cloning' is often used to clone plant genes[7]. Unlike other gene cloning strategies, positional cloning requires no prior knowledge of the gene sequence or product. One crucial step during positional cloning is the production of high-quality genome sequence information spanning the region that contains the gene of interest. Although a reference genome sequence can serve as a 'guide' to narrow down the location of a gene, the gene causing the phenotype of interest is often absent from the reference cultivar[1,4] which means that sequence information from a line that carries the gene of interest is needed. Repeated rounds of bacterial artificial chromosome (BAC) library screening, or chromosome walking, are usually necessary to cover the region of interest with a contiguous sequence[8]. Chromosome walking is particularly tedious in crop species that have large and repeat-rich genomes, such as wheat. The main limitation of BAC clones is that they can only harbor inserts of ~100-200 kb which is why chromosome walking can take a long time. Sequencing and assembly technologies that produce longer sequence scaffolds could prove to be particularly advantageous for positional cloning of plant genes. It has recently been shown that chromosome conformation capture technologies provide powerful tools that enabled the assembly of short sequence reads into long megabase-sized scaffolds in humans and *Drosophila*[9,10].

Leaf rust, caused by the pathogenic fungus *Puccinia triticina* is a widespread and devastating disease of wheat[11] that can be sustainably controlled by exploiting disease resistance that is present in some cultivars of this crop. The disease resistance gene *Lr22a* was crossed into hexaploid bread wheat (*Triticum aestivum*) from its wild relative *Aegilops tauschii* in the 1960s[12]. Following this initial cross, *Lr22a* has subsequently been bred into several Canadian wheat cultivars[13] and *Lr22a*-containing wheat lines have been included in leaf rust surveys worldwide for many years. *Lr22a* confers resistance to a wide range of *P. triticina* isolates [13-16]. The *Lr22a*-mediated resistance is not present in young seedlings (<20 days) but is only visible in wheat plants from ~25 days of age.

First, in order to evaluate the effectiveness of *Lr22a* against Swiss *P. triticina* isolates we inoculated the *Lr22a*-containing backcross line RL6044 ('Thatcher *Lr22a*') and the spring wheat cultivar 'Thatcher' with ten *P. triticina* isolates that were collected in Switzerland. The first leaves of RL6044 developed leaf rust pustules of similar size as the susceptible control 'Thatcher', while we observed complete to moderate resistance on the third leaves of 30-day-old RL6044 plants in comparison to 'Thatcher' (Fig. 1, Supplementary Fig. 1).

Lr22a was previously mapped to the short arm of wheat chromosome 2D using microsatellite analysis of the *Lr22a*-containing wheat line 98B34-T4B[13]. In order to pinpoint *Lr22a* on chromosome 2D we generated a high-resolution mapping population from a cross between the susceptible Swiss spring wheat cultivar 'CH Campala' and an *Lr22a*-containing

backcross line 'CH Campala *Lr22a*'[17] and delimited the gene to a 0.48 cM interval flanked by two microsatellite markers gwm455 and gwm296 (Fig. 2a).

Then, to obtain sequence information for this 0.48 cM interval we used flow cytometry[18] to isolate chromosome 2D from 'CH Campala *Lr22a*', which resulted in ~640 ng high molecular weight DNA of this chromosome. Given the recent examples of long-range scaffolding with the help of chromosome contact maps[9,10], a *de novo* assembly of chromosome 2D was obtained by combining short-read Illumina sequences and proximity ligation of *in vitro* reconstituted chromatin, also known as Chicago[5] (Online Methods). In contrast to the *in vivo* Hi-C method, Chicago has been demonstrated to be more suitable to generate high-quality assemblies from short Illumina reads in vertebrates[5].

The assembly of 'CH Campala *Lr22a*' comprised 10,344 scaffolds with an N50 of 9.76 Mb, that is, half of the chromosome was assembled in scaffolds of 9.76 Mb or more. This N50 is 50-100x longer than a BAC clone and thus each scaffold of this length corresponds to at least 25 rounds of BAC library screening. The longest scaffold was 36.4 Mb and the total assembly was 567 Mb (Supplementary Table 1). The size of the assembly was ~160 Mb shorter than the estimated size of chromosome 2D[19], which was likely due to collapsed high-copy repeats in the assembly (Supplementary Fig. 2). The flanking markers gwm455 and gwm296 were located at a distance of 1.79 Mb on a single scaffold (ScZQ34L_508) of 6.39 Mb in size (Fig. 2b). We used this 'CH Campala *Lr22a*' scaffold to develop additional markers by comparing annotated gene sequences to Illumina reads of the susceptible parent line 'CH Campala'. This allowed us to further reduce the genetic interval to only 0.09 cM (Fig. 2a). The physical distance between the two flanking markers SWSNP4 and SWSNP6 was 438 kb and contained nine genes and two pseudogenes (Fig. 2b). In particular, there was a cluster of two nucleotide binding site-leucine-rich repeat receptor (NLR) encoding genes and two NLR pseudogenes. *NLR1* showed sequence alterations compared to the wild-type 'CH Campala *Lr22a*' allele in five independent EMS mutants that were generated from 'CH Campala *Lr22a*' and that were identified using a phenotypic screen for loss of *Lr22a* resistance. All of the SNPs present in the susceptible mutants are predicted to result in amino acid exchanges or premature stop codons in NLR1 (Fig. 2c, Supplementary Table 2). The sequence of the second full-length *NLR4* gene was identical to the wild-type sequence of 'CH Campala *Lr22a*' in all five susceptible mutants. These results provide evidence that *NLR1* corresponds to the *Lr22a* resistance gene.

To evaluate the overall quality of the 'CH Campala *Lr22a*' assembly, we anchored the scaffolds to a high-resolution genetic map of the wheat D-genome progenitor *Ae. tauschii* that includes 1,326 chromosome 2D-specific single nucleotide polymorphism (SNP) markers[20]. To do this, we performed a BLAST search[21] with the extended sequences of the *Ae. tauschii* SNP markers against the 'CH Campala *Lr22a*' assembly. In total, 1,048

sequences produced BLAST hits that anchored 80 scaffolds (or 521 Mb, which is 92% of the assembly) to the genetic map (Supplementary Table 3). Each of the anchored scaffolds contained an average of 13 SNP markers (range was 1 to 83 markers). We observed a high degree of collinearity between *Ae. tauschii* and the 'CH Campala *Lr22a*' assembly (Supplementary Fig. 3a). Only 62 of the 1,048 genetic markers were non-collinear (mapped to a different location than most markers on the scaffold). Of these, 44 markers were grouped into seven clusters, meaning that at least two markers mapped to a different region on the *Ae. tauschii* genetic map than most of the markers on the scaffold. This might indicate the presence of seven chimeric scaffolds in which two large genomic segments were incorrectly joined. Alternatively, the non-collinear markers might arise by structural variation. The remaining 18 non-collinear markers represented single markers that mapped to a different *Ae. tauschii* position than all others on the respective scaffold; this might by explained by problems in the genetic map of *Ae. tauschii*. SNP markers were perfectly collinear within the *Lr22a* region (Supplementary Fig. 3b).

The predicted coding sequence of *Lr22a* is 2,739 bp, consists of a single exon and translates into a protein of 912 amino acids with an N-terminal coiled-coil domain, a central nucleotide-binding (NB-ARC) domain and a C-terminal leucine-rich repeat domain. In the susceptible parent 'CH Campala', the *NLR1* allele was disrupted by a premature stop codon, whereas the *NLR1* allele in the susceptible wheat line 'Thatcher' was complete and the predicted protein showed 97% amino acid identity to Lr22a (Supplementary Fig. 4). The Lr22a protein only showed weak sequence homology to other cloned wheat NLRs. The closest homolog of Lr22a in Arabidopsis is RPM1, an NLR that confers resistance to the bacterial pathogen *Pseudomonas syringae* (Supplementary Fig. 5). The N-terminal amino acids of RPM1 interact with the RPM1-interacting protein 4 (RIN4), an important regulator of basal defense responses that is targeted by multiple *P. syringae* virulence effectors. Effector-mediated modification of RIN4 is perceived by RPM1, resulting in a hypersensitive response[22,23]. Similarly, it is possible that Lr22a might monitor the status of a basal defense component in wheat. Interestingly, Lr22a contains two amino acids at the N-terminus that are unique compared to the NLR1 protein variants in 25 wheat cultivars without the *Lr22a* resistance (Supplementary Fig. 6).

Several rapid gene cloning methods have been described for wheat[24-26] (Table 1). All of these approaches require the identification of loss-of-function mutants, and some of the methods, i.e. MutRenSeq are only suitable for specific gene classes. However, many agriculturally important genes, for example genes conferring partial disease resistance or abiotic stress tolerance, have 'partial phenotypes' for which the identification of loss-of-function mutants can be challenging. 'Targeted chromosome-based cloning via long-range assembly' offers greater flexibility with respect to gene validation (transformation, haplotype

analysis, TILLING, or genome editing) and since this approach includes the generation of mapping populations, it enables positional cloning of genes with partial phenotypes. Using a cultivar-specific *de novo* assembly, we eliminated the need for chromosome walking. Positional cloning requires high density genetic maps, which are attainable only in distal, telomeric chromosome regions that are characterized by high recombination rates. Pericentromeric and centromeric chromosomal regions show lower recombination rates, which makes the construction of high-density genetic maps challenging. However, long-range scaffolding approaches permit gene cloning even in regions with reduced recombination rates, such as pericentromeric regions and alien introgressions.

For example, for the telomeric region of chromosome arm 2DS, a mapping population of only 400 plants would have been sufficient to reach a 96% probability of finding a target gene and its closest flanking markers on the same sequence scaffold. In pericentromeric regions, where recombination rates are 5-10x lower, a mapping population of 1,200 plants would provide a 90% chance to find a target gene and its closest flanking markers on a single sequence scaffold (Supplementary Fig. 7).

Gene density in wheat and many other grass genomes is highest in distal regions of the chromosome[27,28]. Therefore our approach could find widespread application in cloning most genes in cereals. Critical for the long-range scaffolding of the *Lr22a* region was the amount of DNA required for sequencing because this determined the time needed for chromosome purification. Chicago scaffolding works with small amounts of DNA (~500 ng) and was therefore well-suited to enable a high-quality *de novo* assembly from a flow-sorted chromosome. However, other long-range scaffolding or long-read sequencing technologies that work with small amounts of DNA (<1 µg) such as nanopore sequencing for example might also be used in our gene cloning strategy.

In summary, we report that it is now feasible to develop high-quality *de novo* assemblies from chromosomes of any wheat cultivar. Our approach is applicable for genomic loci of interest and can be applied in species with complex genomes, which should enable cloning of agriculturally important genes. Any species and cultivar from which chromosomes can be flow sorted can be used. To date, flow cytometry has been successfully used in more than 20 plant species, including important crops like maize, wheat, rice, barley, oat, rye, pea, tomato, field bean and chickpea[18].

Methods

Methods and any associated references are available in the online version of the paper.

Accession codes

The 'CH Campala *Lr22a*' scaffolds were deposited at DDBJ/ENA/GenBank under the accession MOLT00000000. The version described in this paper is version MOLT01000000. The *Lr22a* gene sequence was deposited at DDBJ/ENA/GenBank under the accession KY064064. The *NLR1* sequences from 'Thatcher' and 'Campala' have accession numbers KY064065 and KY064066, respectively.

Acknowledgements

We are grateful to the staff at Dovetail Genomics for constructing the 'CH Campala *Lr22a*' scaffolds. We thank M. Karafiátová for supervising chromosome 2D flow sorting and estimation of purity in flow sorted fractions, and Z. Dubská, R. Šperková and J. Weiserová for technical assistance. We also thank B. Senger and L. Luthi for assistance with field experiments and Prof. B. Keller for continuous support. This work was financed by an Ambizione fellowship of the Swiss National Science Foundation. J.V., H.Š. and J.D. were supported by the Ministry of Education, Youth and Sports of the Czech Republic [grant award LO1204 from the National Program of Sustainability I].

Author contributions

A.K.T, T.W., H.Š., J.D. and S.G.K. designed the experiments and wrote the manuscript, A.K.T and S.G.K. performed phenotypic and molecular analyses, H.Š., J.V, and J.D. flow-sorted chromosome 2D and prepared HMW DNA, O.M., C.B. and D.F. developed the 'CH Campala *Lr22a*' backcross line.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Mago, R. *et al.* Major haplotype divergence including multiple germin-like protein genes, at the wheat *Sr2* adult plant stem rust resistance locus. *BMC Plant Biol* **14**, 379 (2014).
2. Jordan, K.W. *et al.* A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol* **16**, 48 (2015).
3. Chia, J.M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* **44**, 803-807 (2012).
4. Rawat, N. *et al*. Wheat *Fhb1* encodes a chimeric lectin with agglutinin domains and a pore-forming toxin-like domain conferring resistance to Fusarium head blight. *Nat Genet* **48**, 1576-1580 (2016).
5. Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26**, 342-350 (2016).
6. FAO. *The state of the world's land and water resources for food and agriculture (SOLAW) - Managing systems at risk*, (Food and Agriculture Organization of the United Nations, Rome and Earthscan, London, 2011).
7. Krattinger, S.G., Wicker, T. & Keller, B. Map-based cloning of genes in Triticeae (wheat and barley). in *Genetics and Genomics of the Triticeae* (eds. Feuillet, C. & Muehlbauer, G.J.) 337-357 (Springer, New York, 2009).
8. Stein, N., Feuillet, C., Wicker, T., Schlagenhauf, E. & Keller, B. Subgenome chromosome walking in wheat: A 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proc Natl Acad Sci USA* **97**, 13436-13441 (2000).
9. Ay, F. & Noble, W.S. Analysis methods for studying the 3D architecture of the genome. *Genome Biol* **16**, 183 (2015).
10. Burton, J.N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-1125 (2013).
11. Kolmer, J. Leaf rust of wheat: Pathogen biology, variation and host resistance. *Forests* **4**, 70-84 (2013).
12. Dyck, P.L. & Kerber, E.R. Inheritance in hexaploid wheat of adult-plant leaf rust resistance derived from *Aegilops squarrosa*. *Can J Genet Cytol* **12**, 175-180 (1970).
13. Hiebert, C.W., Thomas, J.B., Somers, D.J., McCallum, B.D. & Fox, S.L. Microsatellite mapping of adult-plant leaf rust resistance gene *Lr22a* in wheat. *Theor Appl Genet* **115**, 877-884 (2007).
14. Pretorius, Z.A., Rijkenberg, F.H.J. & Wilcoxson, R.D. Characterization of adult-plant resistance to leaf rust of wheat conferred by the gene *Lr22a*. *Plant Dis* **71**, 542-545 (1987).
15. Kolmer, J.A. Virulence in *Puccinia recondita* f. sp. *tritici* isolates from Canada to genes for adult-plant resistance to wheat leaf rust. *Plant Dis* **81**, 267-271 (1997).
16. McCallum, B.D., Seto-Goh, P. & Xue, A. Physiologic specialization of *Puccinia triticina*, the causal agent of wheat leaf rust, in Canada in 2009. *Can J Plant Pathol* **35**, 338-345 (2013).
17. Moullet, O. & Schori, A. Maintaining the efficiency of MAS method in cereals while reducing the costs. *J Plant Breed Genet* **2**, 97-100 (2014).
18. Dolezel, J. *et al.* Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomics* **12**, 397-416 (2012).
19. Safar, J. *et al.* Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res* **129**, 211-223 (2010).
20. Luo, M.C. *et al.* A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc Natl Acad Sci USA* **110**, 7940-7945 (2013).
21. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

22. Belkhadir, Y., Nimchuk, Z., Hubert, D.A., Mackey, D. & Dangl, J.L. Arabidopsis RIN4 negatively regulates disease resistance mediated by RPS2 and RPM1 downstream or independent of the NDR1 signal modulator and is not required for the virulence functions of bacterial type III effectors AvrRpt2 or AvrRpm1. *Plant Cell* **16**, 2822-2835 (2004).
23. Mackey, D., Holt, B.F., Wiig, A. & Dangl, J.L. RIN4 interacts with *Pseudomonas syringae* type III effector molecules and is required for RPM1-mediated resistance in Arabidopsis. *Cell* **108**, 743-754 (2002).
24. Steuernagel, B. *et al.* Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat Biotechnol* **34**, 652-655 (2016).
25. Sanchez-Martin, J. *et al*. Rapid gene isolation in barley and wheat by mutant chromosome sequencing. *Genome Biol* **17**, 221 (2016).
26. Gardiner, L.J. *et al.* Mapping-by-sequencing in complex polyploid genomes using genic sequence capture: a case study to map yellow rust resistance in hexaploid wheat. *Plant J* **87**, 403-419 (2016).
27. Choulet *et al*. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014).
28. Gottlieb *et al.* Insular organization of gene space in grass genomes. *PLoS One* **8**, e54101 (2013).

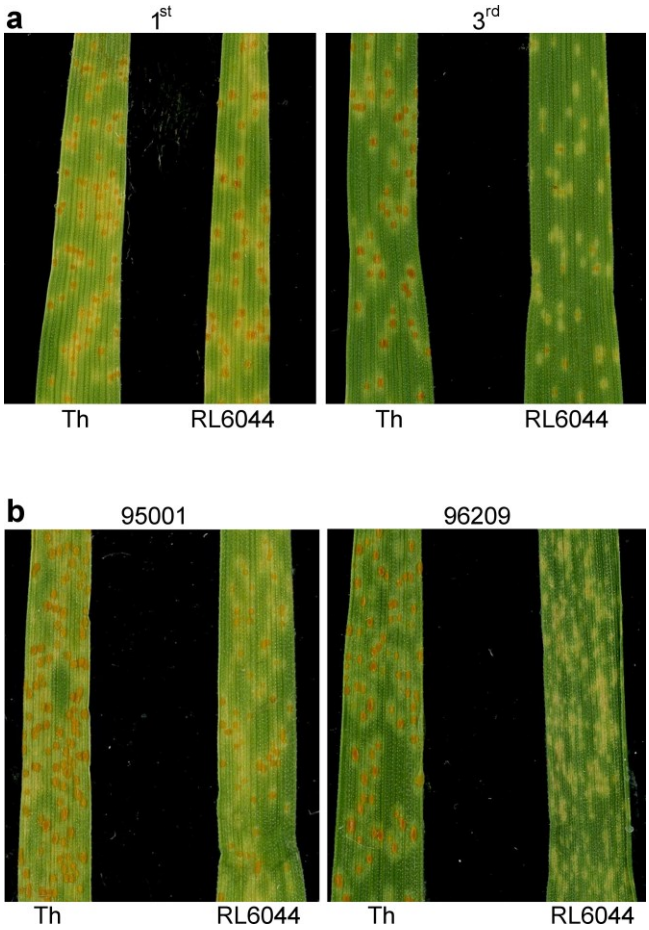

Figure 1. Phenotypic response conferred by the *Lr22a* leaf rust resistance gene. (a) Leaf rust symptoms on first and third leaves of 30-day-old plants of the susceptible cultivar 'Thatcher' (Th) and the *Lr22a*-containing backcross line RL6044 ('Thatcher *Lr22a*'). **(b)** The *Lr22a*-resistance response in RL6044 ranged from partial (left) to complete (right) against different *P. triticina* isolates. Shown here are the two extremes found with *P. triticina* isolates 95001 and 96209.

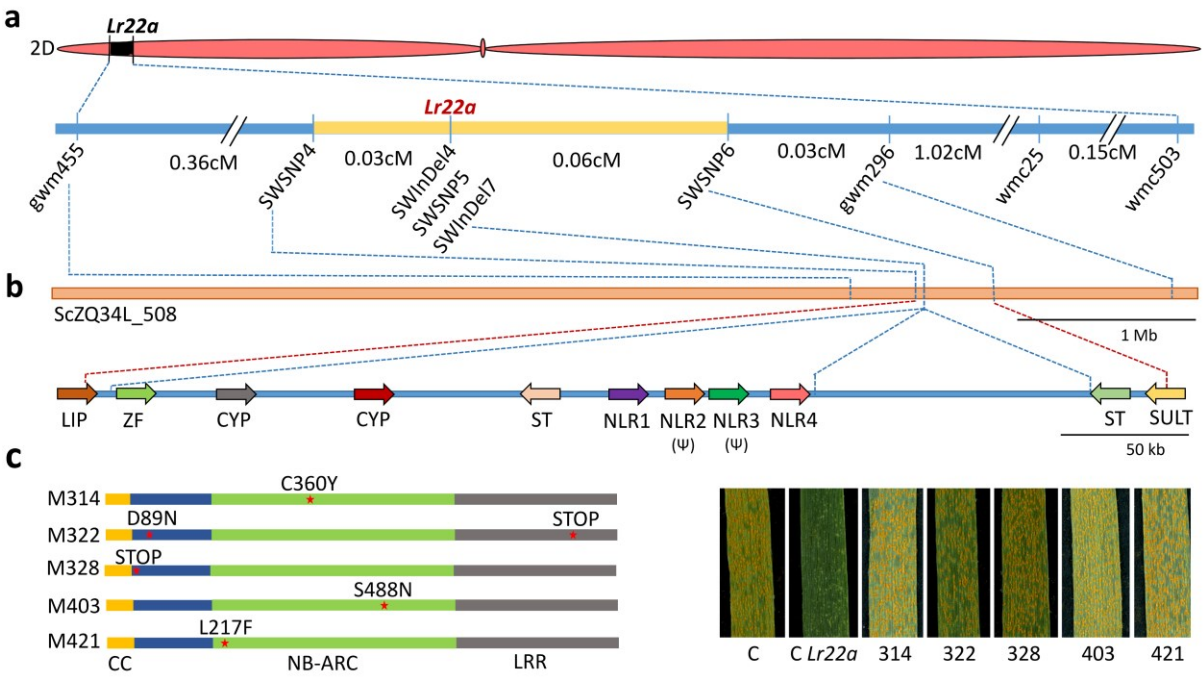

Figure 2. Mapping of the *Lr22a* leaf rust resistance gene. (a) Genetic map of the *Lr22a* region. The target interval between the closest flanking markers SWSNP4 and SWSNP6 is indicated in yellow. **(b)** The physical interval of 'CH Campala *Lr22a*' contained nine candidate genes and two pseudogenes (indicated by Ψ). LIP = lipase, ZF = zinc finger, CYP = cytochrome P450, ST = sugar transporter, NLR = nucleotide binding site–leucine-rich repeat receptor, SULT = sulfotransferase. The 6.39 Mb sequence scaffold ScZQ34L_508 that contained both flanking markers is indicated in orange. **(c)** Five independent EMS mutants that lost the *Lr22a*-resistance had non-synonymous sequence changes in the *NLR1* coding sequence compared to the wild-type allele of 'CH Campala *Lr22a*'. The predicted coiled-coil (CC), nucleotide-binding (NB-ARC) and leucine-rich repeat (LRR) domains of the NLR1 protein are indicated in yellow, green and grey, respectively. The amino acid polymorphisms in comparison to the Lr22a wild-type sequence of 'CH Campala *Lr22a*' are indicated by red asterisks. C='CH Campala', C *Lr22a* = 'CH Campala *Lr22a*'.

Table 1. Comparison of different gene isolation approaches

	MutChromSeq[25]	MutRenSeq[24]	Mapping-by-sequencing[26]	Positional cloning by chromosome walking	Targeted-chromosome-based-cloning via long-range assembly
Dependence on enrichment library	No	Yes	Yes	No	No
Dependence on the identification of loss-of-function mutants	Yes	Yes	Yes	No	No
Dependence on reference sequence	No	Depends on reference gene annotation for enrichment library	Depends on reference gene annotation for enrichment library	No	No
Speed / cost-effectiveness	Very rapid, cost-effective	Very rapid, cost-effective	Very rapid, cost-effective	Very slow, expensive	Rapid, cost-effective
Major limitations	Depends on the identification of loss-of-function mutants, no backup if mutants cannot be identified	Only allows identification of NLRs, depends on enrichment library, depends on the identification of loss-of-function mutants, no backup if gene of interest does not encode a NLR	Depends on enrichment library or a high-quality reference sequence, depends on the identification of loss-of-function mutants	Very slow, a cultivar-specific BAC library is often necessary, depends on recombination	Partially depends on recombination, but also works in chromosomal regions with reduced recombination rates
Best suited for	Isolation of genes with strong phenotypes	Isolation of NLRs with strong phenotypes	Isolation of genes with strong phenotypes	Any gene, also suitable for genes with partial phenotypes and adult plant phenotypes	Any gene, also suitable for genes with partial phenotypes and adult plant phenotypes

Online Methods

Plant material

The *Lr22a*-containing wheat lines RL6044[12] (Thatcher*7//tetra-Canthatch/RL5271) and 'CH Campala *Lr22a*'[17] (CH Campala*6/AC Minto) were used in this study. A bi-parental mapping population consisting of 1,656 F2 plants was derived from a cross between 'CH Campala *Lr22a*' and the susceptible near isogenic Swiss spring wheat cultivar 'CH Campala'. DNA was extracted from leaf tissues using a CTAB extraction protocol[29]. A total of 1,656 F2 plants were screened for recombination between SSR markers gwm455 and wmc503[13]. Polymerase chain reaction (PCR) products were separated on polyacrylamide gel using a LI-COR® DNA Sequencer 4200. In total, 54 recombinant F2 plants were identified showing 55 recombination events between the two markers. F3 families of recombinant F2 plants were phenotyped in the field and growth cabinets. F3 families were classified as uniform susceptible, uniform resistant or segregating based on a comparison to the two parents. In addition, homozygous recombinant F4 families were selected and re-phenotyped in growth cabinets. Field infections were done as described previously[30]. For the infection assays in growth cabinets, five seeds per family were sown in two replicates in soil in 1.5 l pots. After treatment with 4 ml/l growth inhibitor (Cycocel® Extra, Omya AG, Oftringen, Switzerland) and 2-3 ml/l fertilizer (Wuxal® Profi, Maag Garden, Syngenta, Dusseldorf, Germany) plants were grown at 20°C and a 16 h photoperiod (450 µmol m^{-2} s^{-1}) followed by 8 h at 16ºC without light and a relative humidity of 70%. Plants were inoculated with *P. triticina* isolate 90035 suspended in oil (Fluorinert™ FC-43, 3M Electronics, Zwijndrecht, Belgium) when they were 20-25 days old. After the inoculation, plants were kept in the dark for 24 h under a plastic tent to maintain high humidity and then shifted back to normal growth conditions. Disease symptoms were assessed 10 days after inoculation.

EMS mutagenesis and identification of *Lr22a* mutants

Ethyl methanesulfonate (EMS) mutagenesis was performed as described previously[31]. In a preliminary experiment, 0.35% EMS was identified as the concentration that resulted in 50% seedling mortality. Then, 1,100 seeds of 'CH Campala *Lr22a*' were soaked in water at 4ºC for 16 h and then the treatment with 0.35% EMS was done for 16 h at room temperature with constant shaking at 150 rpm. Treated seeds were washed in tap water and plants were advanced to M2 generation in the glasshouse. Seeds of 685 M1 plants were harvested and the respective M2 families were screened for susceptibility with the *P. triticina* isolate 90035 as described above. Out of this screen, five susceptible mutants derived from different M2 families were identified and validated in the M3 generation. All susceptible mutants identified in this screen carried sequence polymorphisms in NLR1.

Flow sorting of chromosome 2D and preparation of DNA samples

Chromosome 2D was purified by flow cytometric sorting as described earlier[32,33] with modifications. Briefly, suspensions of intact mitotic metaphase chromosomes were prepared from synchronized root tip meristem cells. Before flow cytometry, GAA microsatellites were labelled on chromosomes in suspension by FITC following a previously described protocol[34] and chromosomal DNA was stained by DAPI (4´,6-diamidino 2-phenylindole). Chromosome samples were analyzed at rates of 1,500 – 2,000 particles / sec on a BD FACSAria SORP flow cytometer (Becton Dickinson Immunocytometry Systems, San Jose, USA) and bivariate flow karyotypes FITC vs. DAPI fluorescence were acquired. Sort windows delimiting the population of chromosome 2D were set on dotplots FITC vs. DAPI (Supplementary Fig. 8), and chromosome 2D was sorted at rates of 15 - 20 / sec. The identity of flow-sorted chromosomes and contamination by other chromosomes were checked microscopically using fluorescence *in situ* hybridization (FISH) as descried previously[35] using probes for GAA microsatellites and *Afa* family repeat.

For shotgun sequencing, DNA of chromosome 2D was amplified by multiple displacement amplification (MDA) as described previously[36]. In total, 30,000 copies of chromosome 2D were flow sorted from each line. The purity of the sorted fraction was 94%. The chromosomes were treated with proteinase K and the purified DNA was amplified using an Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Chalfont St. Giles, UK). Three independent MDA products from each sorted chromosome fraction were pooled into one sample to reduce amplification bias.

For long-range assembly, high molecular weight (HMW) DNA was prepared from flow-sorted chromosome 2D of 'CH Campala *Lr22a*' following a previously described protocol[37] with modifications. A total of 1,5 million copies of chromosome 2D were flow sorted with a purity of 97% and embedded in six agarose miniplugs with a total volume of 100 µl. Plugs were then incubated in proteinase K. The miniplugs were washed six times in TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0), melted for 5 min at 73°C and solubilized with 0.8 U GELase (Epicentre, Madison, USA) for 45 min. The released DNA underwent 60 minutes of drop dialysis (Merck Millipore, Billerica, USA) against TE buffer. Purification and concentration was performed using a Vivacon® 500 centrifugal concentrator (100,000 Dalton MWCO, Sartorius, Goettingen, Germany). The HMW DNA was partially fragmented by pipetting and vortexing to facilitate concentration measurement.

Establishment of long-range assembly from 'CH Campala *Lr22a*'

Chromosome 2D shotgun sequencing, Chicago sequencing and scaffolding was performed by Dovetail Genomics (Santa Cruz, CA). A Chicago library was prepared as described

previously[5]. Briefly, 250 ng of chromosome 2D HMW DNA (mean fragment length ~100 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with *Mbo*I, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and a sequencing library was generated using NEBNext® Ultra™ enzymes (New England BioLabs) and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment. The library was then sequenced on an Illumina HiSeq 2500 (rapid run mode) to produce 145 million 150 bp paired-end reads, which provided 30x physical coverage of the chromosome (1-50 kb pairs).

De novo chromosome 2D assembly was constructed using sequence data from three paired-end libraries, two prepared from 50 ng of chromosomal DNA with a mean insert size of 205 bp and one prepared from 150 ng of chromosomal DNA with a mean insert size of 450 bp. The libraries were sequenced on an Illumina HiSeq 2500 (rapid run mode) to produce a total of 709 million 150 bp paired-end reads (312 million from the shorter insert libraries and 397 million from the longer insert library). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic[38]. *De novo* assembly was performed using Meraculous 2 (2.2.2.3)[39] with a kmer size of 109.

The input *de novo* assembly, shotgun reads, and Chicago library reads were used as input data for HiRise, a software pipeline designed specifically for using Chicago data to scaffold genome assemblies[5]. Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify putative misjoins and to score prospective joins. After scaffolding, the shotgun sequences were used to close gaps between contigs. To generate a pseudomolecule, the extended sequences of 1,326 chromosome 2D-specific SNPs mapped to the *Ae. tauschii* AL8/78 genetic map[20] were used to perform a BLAST search against the 10,344 'CH Campala *Lr22a*' scaffolds using an in-house script.

Marker Development

SSR markers gwm455, gwm296, wmc503 and wmc25 were previously reported to be linked to *Lr22a*[13]. For the development of additional markers, a *de novo* Illumina sequence assembly was developed from DNA amplified from flow-sorted 2D chromosomes of 'CH Campala' and 'CH Campala *Lr22a*'. DNA from chromosome 2D of each parent were multiplexed and sequenced on one lane of an Illumina HiSeq 2500 with 125 bp paired-end

reads. The sequencing was performed at the Functional Genomics Center Zurich, Switzerland. The reads were used for a *de novo* assembly using CLC Main Workbench 7 (Qiagen) with standard parameters and a minimum contig length of 500 bp. For 'CH Campala', 84 million reads were obtained and assembled into 57,314 contigs with a total size of 123 Mb and a scaffold N50 of 3.8 kb. For 'CH Campala *Lr22a'*, 139 million reads were obtained that were assembled into 71,348 contigs with a total size of 159 Mb and a scaffold N50 of 3.9 kb. Illumina contigs were filtered for contigs containing genes by performing a BLAST search[21] against the *Brachypodium distachyon* coding sequence database[40]. Gene-containing contigs were used for the discovery of SNPs and insertions/deletions (InDels) based on a previous protocol[41] with minor modifications. The sequences of the *Lr22a* flanking SSRs gwm455 and wmc25 were anchored to the genetic map of *Ae. tauschii* AL8/78[20] by performing a BLAST search against the *Ae. tauschii* BAC scaffolds[20,42] (http://aegilops.wheat.ucdavis.edu/ATGSP/). This resulted in the identification of two scaffolds, 4242.1 (gwm455) and 4531.6 (wmc25). A second BLAST search with the identified scaffolds against the extended sequences of the *Ae. tauschii* SNP markers (http://probes.pw.usda.gov/WheatDMarker/) identified the chromosome 2D-specific markers AT2D1039 and AT2D1040 (gwm455) and AT2D1053 (wmc25) that were located at cM positions 25.59 – 28.502 on the *Ae. tauschii* genetic map. The extended sequences of markers mapped between AT2D1039 and AT2D1053 were then used to perform a BLAST search against the *Ae. tauschii* BAC scaffolds. The *Ae. tauschii* BAC scaffolds were used to identify the corresponding sequences in the gene-containing contigs of 'CH Campala' and 'CH Campala *Lr22a'*. The identified contigs of the two wheat lines were aligned using Clustal Omega[43] and locus-specific PCR probes spanning polymorphisms between 'CH Campala' and 'CH Campala *Lr22a*' were developed and sequenced on the recombinants of the fine-mapping population. This resulted in the development of two markers, SWSNP5 and SWInDel4 (Supplementary Table 4). Similarly, the 'CH Campala *Lr22a*' Chicago assembly was used to develop additional markers. Scaffold ScZQ34L_508 was annotated using the *B. distachyon* coding sequence database (https://phytozome.jgi.doe.gov/pz/portal.html). The Illumina contigs of 'CH Campala' were mapped against the annotated genes using BLAST and SNPs and InDels were identified as described above. This resulted in the development of three additional markers, SWSNP4, SWSNP6 and SWInDel7. For the amplification of the *Lr22a* gene, specific primers (LRR1-F3 and LRR1-R4) were designed from the 5' and 3' UTR and amplified using the Kapa HiFi HotStart PCR kit (KapaBiosystems) according to the manufacturer's protocol. The amplicon was sequenced using eight internal primers (Supplementary Table 4).

Lr22a protein domain prediction

The predicted Lr22a protein sequences from RL6044 and corresponding NLR1 protein version from 'Thatcher' were aligned using the online Clustal Omega[43]. Different domains of the NLR were identified based the homology to the annotated RPM1 protein[44]. The most probable LRR motifs were predicted using the LRR conservation mapping tool v2.0 (www.plantpath.wisc.edu/RCM)[45].

Statistical methods

A phylogenetic tree of Lr22a and known wheat resistance proteins was made using the PROTPARS tool of the PHYLIP package with 100 bootstrap replicates[46]. The amino acid sequences of known wheat NLRs were downloaded from the NCBI repository. Amino acid sequences of the LRRs were aligned using ClustalX 2.1 using a gap opening penalty of 10 and a gap extension penalty of 0.2.

Simulation of recombination frequencies and population sizes

The goal of this simulation was to calculate the probabilities of finding a target gene and its closest flanking markers on a single sequence scaffold using different sizes of mapping populations. Recombination frequencies were derived from combining the genetic mapping data from *Ae. tauschii*[20] and the physical sizes of the 80 'CH Campala *Lr22a*' scaffolds that were anchored to the genetic map (Supplementary Table 3). Local recombination frequencies (in Mb/cM) along chromosome 2D were calculated in a sliding window averaging ratios of physical to genetic distance over 50 genetic markers. Based on the resulting recombination frequency graph, we divided the chromosome into two telomeric, two pericentromeric and one centromeric bin (Supplementary Fig. 7a). Simulations were run for the two telomeric bins separately, because recombination frequencies on the short and long arm telomeric bins differed by a factor of 2.3 (median 1.2 Mb/cM for 2DS vs. 2.75 Mb/cM for 2DL, Supplementary Fig. 7a). Data for pericentromeric bins were compiled resulting in a median recombination frequency of 9.1 Mb/cM (Supplementary Fig. 7a). The simulation used real-life set of sizes of the 80 sequence scaffolds. These were randomly picked until the cumulative size had reached the size of the respective chromosome bin. Then, the target gene was positioned randomly inside the bin. Next, the recombination breakpoints were distributed randomly across the chromosome segments (assuming recombination frequency to be evenly distributed along the bin). The number of recombination breakpoints was determined by population size and recombination frequency for the respective bin. Finally, the software tested whether the target gene was flanked by two recombination breakpoints on the same sequence scaffold. The sizes of tested mapping populations ranged from 50 – 2,000 individuals, increasing the population size in steps of 50. The simulation was repeated

10,000 times for each population size, which provided the probabilities of the gene being flanked on both sides by genetic markers on the same sequence scaffold for different sizes of mapping populations (Supplementary Fig. 7b). All original Perl scripts used for calculations of recombination frequencies and simulations are available upon request.

References

29. Stein, N., Herren, G. & Keller, B. A new DNA extraction method for high-throughput marker analysis in a large-genome species such as *Triticum aestivum*. *Plant Breeding* **120**, 354-356 (2001).
30. Singla, J. *et al.* Characterization of *Lr75*: a partial, broad-spectrum leaf rust resistance gene in wheat. *Theor Appl Genet* **130**, 1-12 (2017).
31. Periyannan, S. *et al.* The gene *Sr33*, an ortholog of barley *Mla* genes, encodes resistance to wheat stem rust race Ug99. *Science* **341**, 786-788 (2013).
32. Vrána, J. *et al.* Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* **156**, 2033-2041 (2000).
33. Kubaláková, M., Vrána, J., Číhalíková, J., Šimková, H. & Doležel, J. Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* **104**, 1362-1372 (2002).
34. Giorgi, D. *et al.* FISHIS: fluorescence *in situ* hybridization in suspension and chromosome flow sorting made easy. *PLoS One* **8**, e57994 (2013).
35. Kubaláková, M. *et al.* Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome* **46**, 893-905 (2003).
36. Šimková, H. *et al.* Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* **9**, 294 (2008).
37. Šimková, H., Číhalíková, J., Vrána, J., Lysák, M.A. & Doležel, J. Preparation of HMW DNA from plant nuclei and chromosomes isolated from root tips. *Biol Plantarum* **46**, 369-373 (2003).
38. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
39. Chapman, J.A. *et al.* Meraculous: *de novo* genome assembly with short paired-end reads. *PLoS One* **6**, e23501 (2011).
40. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763-768 (2010).
41. Shatalina, M. *et al.* Genotype-specific SNP map based on whole chromosome 3B sequence information from wheat cultivars Arina and Forno. *Plant Biotechnol J* **11**, 23-32 (2013).
42. Jia, J. *et al. Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91-95 (2013).
43. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
44. Gao, Z., Chung, E.H., Eitas, T.K. & Dangl, J.L. Plant intracellular innate immune receptor Resistance to *Pseudomonas syringae* pv. *maculicola* 1 (RPM1) is activated at, and functions on, the plasma membrane. *Proc Natl Acad Sci USA* **108**, 7619-7624 (2011).
45. Helft, L. *et al.* LRR conservation mapping to predict functional sites within protein leucine-rich repeat domains. *PLoS One* **6**, e21614 (2011).
46. Retief, J.D. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* **132**, 243-258 (2000).